Formerly
DOE-HDBK-1205-97

DESIGN, DEVELOPMENT, AND IMPLEMENTATION OF EXAMINATIONS

U.S. Department of Energy
Washington, D.C. 20585

FOREWORD

This document provides information to DOE staff and contractors that can be used by training staffs (e.g., instructors, designers, developers and managers) and others for the construction and administration of written, oral, and performance examinations. This document has been prepared on the basis of training programs used at various DOE nuclear facilities. Users are not obligated to adopt any part of this document; rather, they can selectively use the information to establish or improve facility training programs as applicable. This document was based upon DOE Handbook, *Guide to Good Practices for the Design, Development, and Implementation of Examinations*.

Beneficial comments (recommendations, additions, and deletions) and any pertinent data that may be of use in improving this document should be addressed in the Comments Section of this forum.

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Purpose

The purpose of this document is to provide information to training personnel in the broad areas of design, development, and implementation of examinations.

1.2 Discussion

Nuclear facilities spend a significant amount of training resources testing trainees. Tests are used for employee selection, qualification, requalification, certification and recertification, and promotion. Ineffective testing procedures, or inappropriate interpretation of test results, can have significant effects on both human performance and facility operations. Test development requires unique skills, and as with any skill, training and experience are needed to develop the skills. Test development, test use, test result interpretation, and test refinement, like all other aspects of the systematic approach to training, should be part of an ongoing, systematic process.

For some users this document will provide a review of ideas and principles with which they are already familiar; for others it will present new concepts. While not intended to provide in-depth coverage of test theory design and development, it should provide developers, instructors, and evaluators with a foundation on which to develop sound examinations.

2. THE PURPOSES OF TESTING

There are several reasons for using tests in job and training environments. These include:

- Trainee assessment
- Trainee selection and placement
- Trainee motivation
- Instructional improvement
- Program evaluation
- Testing as a teaching instrument.

These reasons each have their benefits and are applicable to the development and conduct of DOE training programs. In a program based on the systematic approach to training, tests are normally designed and developed for the purpose of trainee assessment. However, the other reasons listed can also be achieved by analyzing and interpreting the test results.

2.1 Trainee Assessment

When designing a test, the purpose of the test should always be to evaluate which learning objectives have been met and therefore the knowledge acquired, or which tasks or partial tasks a trainee is qualified to perform. However, the results of the test may be used to tell us more than that when combined with other evaluation results. For example, results may indicate that:

- the trainee qualifies for advanced placement or exception
- the material on a particular subject needs upgrading
- a particular test question is poorly worded.

When properly developed and conducted, testing should provide a valid and reliable indicator of trainee performance. Whether written, oral, or performance, tests provide the most complete and efficient method of collecting and documenting data on trainee performance. Observation interviews and other applied research methods can also offer a significant amount of information.

The most common test type for the performance-based training environment is the performance test. These tests are normally developed by the facility training department and are used to qualify, re-qualify, and test the progress of trainees for a specific task,

job, or position. A pass/fail system is based on a pre-established cutoff score. Most tests of this type are not designed to identify unique trainee strengths or weaknesses, but can be used for this purpose.

2.2 Trainee Selection and Placement

Tests are useful for trainee selection and placement. Entrance tests are sometimes used as a basis for the initial decision for hiring an employee or for waiving specific job training requirements after that employee begins work. Test scores may also indicate the need for remedial training. Many facilities use aptitude tests for placement of trainees in training programs. Some use interest inventories and psychological evaluations to aid in job placement. If a facility uses placement tests, the facility management, personnel department, and training staff should make placement decisions based, in part upon test scores.

2.3 Trainee Motivation

Tests are powerful motivators. Trainee study habits can be affected by test schedules. When tests are given on a daily or weekly basis, trainees may study more in anticipation of those tests. Likewise, when there is only an end-of-course test, many trainees may postpone studying until just before the test. Trainees are also generally motivated by the feedback that a test score provides. Low test scores could raise trainees' anxiety levels, which if properly channeled, can result in increased concentration and study.

2.4 Instructional Improvement

Test results can provide constructive feedback regarding the effectiveness of a training program. If an instructor fails to adequately cover a topic in a classroom presentation or a laboratory exercise, lower test results may reveal the omission. Uniformly high scores for a topic or subject area may indicate that instruction was effective, and can signify the readiness of the trainees for more detailed material or for the next step in the training program. Conversely, low scores may indicate a need for improvement in instruction or teaching material, or that more instruction time is needed. For example, if a significant number of trainees missed questions based on one learning objective, training may be found to be inadequate for that objective.

2.5 Program Evaluation

Trainee test scores should be combined and analyzed to obtain course or program performance information. This information can be valuable in assessing program strengths and weaknesses. To maximize the usefulness of the test data, systematic reviews should be conducted. Programmatic information can be obtained by analyzing

and interpreting the results of tests, and then comparing that data with information acquired from instructor, supervisor, and trainee questionnaires. When combined, these sources can form a composite picture of program strengths and weaknesses, and appropriate actions can be taken to correct deficiencies.

2.6 An Instrument to Provide Feedback

Instructors who view testing as only an evaluation tool often overlook the opportunity to use testing as a learning tool. An application of testing as teaching can be seen in the On-the-Job Training (OJT) process. In this activity, the trainees perform tasks under the supervision of a subject matter expert (SME). If the trainees perform properly, the performances are acknowledged; if not, the trainees are given immediate feedback on what errors were made and the proper steps needed to correct them. When this occurs during the training process, the trainees may have the opportunity to make the corrections at once. Testing can also provide effective feedback in the classroom, especially when test results are reviewed with the trainees. An open discussion of incorrect answers and why the wrong answers were selected can be very beneficial for both trainee and instructor.

3. BASIS OF THE TEST

Test design and development should not be viewed as a strictly defined mechanical process with blind application of testing principles. Rather, a good test is the direct result of the implementation of testing principles. The test developer should be knowledgeable of good testing principles, the subject matter, its significance, and the most appropriate training setting and method for the material.

3.1 Analysis Prior to Testing

Specific areas to be tested should be proven important to job performance. Proper analysis of the job or task for which the trainee is being trained provides direction for the entire training program. Tasks required for competent job performance are identified, documented, and included in the training program as a result of a job analysis. Learning objectives that identify training content and define satisfactory performance are derived from these tasks. Effective testing requires learning objectives to be carefully selected and classified prior to test development.

Detailed discussions of three types of analyses (needs, job, and task) are found in the *DOE Training Program Handbook: A Systematic Approach to Training*. Alternative methods for analysis are discussed in the DOE Handbook *Alternative Systematic Approaches to Training.* DOE nuclear facilities should perform plant-specific analyses that provide detailed bases for their training programs. These analyses should be conducted by personnel who have been trained to conduct analyses of training requirements.

3.2 Learning Objectives

Learning objectives identify the knowledge and skills that are necessary to perform the job or task. A properly designed learning objective will allow the test developer to determine the specific aspects of the knowledge, skills, and ability to be measured by the test item.

Learning objectives also provide the conditions under which the test will take place and a standard against which test items are judged.

Along with learning objectives, the test developer should review any available supporting instructional materials and other facility reference material to assist in test development. For a more detailed discussion on learning objectives refer to *Developing Learning Objectives* in the SAT portion of this forum.

3.3 Test Banks

Facility training departments should develop and maintain test banks. These banks should consist of previously used tests, answer keys, and test items. Not only do these test banks save a great deal of time, but the resulting tests are significantly improved because of any modifications made following the use of each test. Training programs should include such a test bank and instructors should collect test analysis information each time a test is used. Since facility training organizations may provide training by program area using several instructors, it is important that the test bank concept be applied at the program level. In this way, the size, scope, and uniformity of the testing process will be improved.

The widespread use of computers and data-base software has added significantly to the capabilities and flexibility of such systems. For example, multiple versions of a test may be produced to increase test security during administration. There is a large amount of written test generation and records maintenance software systems available to increase the ease and efficiency of test development and administration. These systems provide an effective tool for test item evaluation and improvement.

Test Bank Establishment Considerations

The following should be considered when establishing test banks:

- The scope of the bank
- Effective security controls for computerized test banks
- An ongoing program for test and test item analysis
- The use of machine-scored answer sheets as appropriate
- Clear guidelines and procedures
- A test outline or test specifications
- A test item numbering system. The following is a list of potential test item identifiers:
    - Program
    - Procedure number
    - Lesson plan
    - Learning objective
    - Test item format
    - Test item level
    - Point value

  - Date test item is generated
  - Test item generated
  - Dates test items are used on tests.
- An ongoing program for test item review, replacement, and new test items
- Sharing information with other facilities.

3.4 Selection of Test Format

There is no single test format for all situations. A format appropriate in one environment may be less appropriate in another. Each format has its advantages and disadvantages. Test quality depends on the quality of the learning objectives and the consistency between these objectives and the test items. The test developer may consider the following factors when developing tests.

Facilities Available

If time permits, the actual job environment may be used to perform the test. Ideally, training environments are divided into "classroom," "laboratory," "simulator," and "OJT," with each environment using the most appropriate test format.

Number of Trainees

The number of trainees that take a test can impact on the format chosen for the test. A key advantage of certain formats is quick scoring. If a test is used for a large number of people, this may be the best choice. However, quality should never be sacrificed for quantity.

Time

Essay tests generally require more time to administer and score. Essay tests may require an hour to administer four questions, while four multiple choice questions can typically be completed in a few minutes. The length of a written test should not exceed the number of test items which could be answered in two hours by the average trainee. This may require assembling several tests for a given instructional area. Time is also a factor in the administration of performance tests. It can easily take several hours to set up and administer a performance test on a simulator or in an on-the-job location.

Before tests can be developed, the appropriate test format should be selected. There are three basic formats:

- Written tests
- Oral question tests
- Performance tests.

4. WRITTEN AND ORAL TESTS

When written and oral tests are designed and developed, several decisions should be made early in the process. These decisions include:

- Specific learning objectives to be tested
- Format for the test
- Amount of emphasis each test item receives
- Number of items on the test
- Time allowed for the test
- Statistical properties of test items such as difficulty and discrimination, where appropriate.

4.1 Open Reference Test

Open reference or open book testing is when the reference, or a sufficient subset of the reference, is provided to the trainee during administration of a test. The test developer should determine which references and their applications are necessary after reviewing the learning objectives and the test specifications. While the open reference test is essentially no different than other written tests, there are several points to consider when using this method.

Ensure that the trainees are aware the test will include an open reference section as well as what references will be made available to them. Listing specific references may not be necessary; however, the references and job aids should be made available during testing consistent with the conditions stated in the learning objectives. This step is important because trainees need to know what will be expected during testing (i.e., using references rather than memorizing them).

Administer all closed reference test items separately and prior to the open-reference test section. This ensures that the trainees do not find so-called "giveaway" answers in the references.

Allow sufficient time for the trainees to complete the open reference questions. The more familiar trainees are with the references, the faster they can complete the items. However, be cautious the test does not become a time test. Unless time is a crucial factor in the task, it should not be made a part of the test.

4.2 Test Specifications

Test specifications are a blueprint, or plan, that clearly defines the scope and content of the test. It is the documentation for the decisions made in the initial planning stages. Just as it is important to develop learning objectives before instruction is planned, it is necessary to develop test specifications prior to test construction.

The development of test specifications is a vital step in the testing process. Test specifications provide two important checks on the entire test mechanism. They are:

- An explicit, documented link between each test item and a learning objective that is verified to be relevant, important, and based on the task
- Consistency in the way tests are developed at a facility.

Consistency will assist in reducing biases in test content due to instructor likes and dislikes or the changing of personnel at the facility. The process ensures all decisions for job placement are based on trainee performance on the same body of knowledge and ability, even though specific topics covered on individual tests may differ.

Developing Test Specifications

Since learning objectives complete with action statements, conditions, and standards already exist, the major portion of test planning is accomplished. What remains is to determine which objectives will be covered in the test, how many items will be included, and which test items are of relative importance. When developing test specifications for exams, it is important to recognize that the knowledge and skills for all learning objectives should be tested at some point in the training.

Table 1 shows test specifications developed from a list of learning objectives. The objective statements indicate the type and level of performance expected of the trainee. The instructor should select the objectives that will be tested on a given exam and establish the relative emphasis each learning objective receives.

Table 1. Test Specifications.

Objectives for Training	Testing Emphasis (item weight %)	Objectives to be Included in Test
I. Area A		
Objective-1	5	Yes
Objective-2	10	Yes
Objective-3	0	No
Objective-4	5	Yes

II. Area B		
Objective-1	10	Yes
Objective-2	0	No
Objective-3	2	Yes
Objective-4	0	No
III. Area C		
Objective-1	5	Yes
Objective-2	10	Yes
Objective-3	2	Yes

As Table 1 shows, Objective III.2 is given twice as much weight on the test as III.1 and five times as much weight as III.3. These different weights are based on the objectives' comparable importance to success in job performance and should reflect the relative time spent on the objectives during the course of the training program. There are no pre-established rules for determining the specific weight assigned to various cells of test specifications. However, the objectives that represent task elements that are critical to the successful accomplishment of the task need to be tested and those test items cannot be missed. The test developer should obtain input from other trainers, from subject matter experts (SMEs), and from facility operations management and supplement this with his/her own prior experience. Trainees will expect the testing emphasis to be comparable to the emphasis stressed during training, and this should be the case. Learning objectives can be assigned greater emphasis by increasing the number of test questions for those objectives.

Table 1 further shows that Objectives I.3, II.2, and II.4 do not appear in this test. This is because they were covered in previous tests, will be covered in later tests, or can be tested in conjunction with other objectives. All learning objectives should be tested at some point during the training (but not necessarily on the final examination) or consideration should be given to their importance to the overall objective of the specific subject area. That is, if it is not considered important enough to be tested, it most likely is not important enough to be included in the training program.

The completed test specifications (test outline or sample plan) provide the test developer with a list of learning objectives upon which to base test questions.

4.3 Test Construction

The actual test is constructed following test design and test item development. Test construction requires the test developer to establish the test layout, assemble the test items, prepare the answer key, and write test directions. Test construction should be completed before implementing the training program.

Test developers should construct tests to some predetermined standardized appearance. The layout and source for this appearance is not as important as maintaining consistency for all of the facility's tests. This consistency of appearance has several advantages. One advantage is minimizing trainee stress by providing a layout trainees are familiar with; test day is an inappropriate time for introduction of a new test format or layout. Another advantage is improved reliability. Inherent reliability is based on the consistency or dependability of test results. Trainees should be tested in a similar manner from test to test. This involves similar test lengths, consistent grading and scoring, use of the same construction guidelines, consistent coverage of topics, familiar test item formats, etc.

Facilities should have procedures in place that establish the format and layout of their training program tests. Tests are assembled using the following general guidelines.

- Select the appropriate test items based on the test specifications.
- Prepare the test key when the test is constructed.
- Indicate and be consistent with point allocations for each answer in regard to the importance of the learning objective that the test item is testing.
- Assign the number of questions per content area that reflects the appropriate emphasis.
- Change the tests' content from one test to the next so they are not compromised.

4.4 Test Layout and Assembly

The test should be assembled in a logical and easily understood format and should follow conventional rules of order for the test items.

Written tests should include typed or printed test items (no handwritten tests) and should be reproduced so each trainee has a test. Writing the questions on the board or stating the questions orally invites misunderstanding. An oral examination is not meant to be a written test given orally; rather, it is a unique situation requiring two-way communication.

The test should be clearly labeled. The course, test title, associated unit of study, administration date, and test form should be stated on the test. If the test is to have trainee responses written on it, put this identifying information on a cover page where the trainee's name, employee number, or other required information is entered. The preferred arrangement of test items is to group:

- All items using a common body of supporting information (e.g., diagram, table, or scenario) even if test item formats are mixed

- All items of the same format
- All items dealing with the same learning objective
- Items from least to most difficult.

Some tests consist of only one format, but most tests contain a variety of formats. While using only one format has the advantage of simplicity and clarity in giving only one set of directions, it is more difficult and time consuming for the test developer to force all questions into one format. There is nothing wrong with a variety of formats; however, to keep the test responses ordered from simple to complex, the following order of test items is suggested:

- Multiple choice items
- Matching items
- Short answer items
- Essay questions.

When a diagram, drawing, or block of information is used with a test item or items, place it above or below the test question if possible. If it is too large to go on the same page as the test item, it should be attached as the next page in the test so the trainee does not have to search for it. The test item should state the location of the diagram, drawing, etc., if not on the same page. Avoid splitting a test item's material between two pages, but if one is split, present all of the item alternatives on the same page. Keep matching items together on the same page.

Consideration should be given to placing only one question per test page. This minimizes the administrative burden on the trainee, improves test clarity, and reduces the chances of the trainee inadvertently failing to answer a question.

4.5 Written Test Administration

Test administration has an important effect on the usefulness of test results and requires control. The instructor should ensure that a suitable environment is established, consistent and clear test directions are given, and proper supervision is present for the entire test.

Establish Environment

Effective testing environments require attention to the physical qualities of the test setting and to the trainees' emotional climate. High noise levels, poor lighting, lack of ventilation, excessive heat or cold and frequent interruptions will lower trainee test performance. The instructor should optimize, to the extent possible, the conditions for

testing. This may be as simple as scheduling testing in the morning if the classroom becomes too hot in the afternoon.

While most instructors are aware of the physical testing environment, many do not give sufficient consideration to the emotional environment they establish. The testing environment should be conducive to effecting testing. Making the purpose of the test clear and emphasizing the need for accurate test results can create a good emotional climate, which is important in building motivation, reducing anxiety, and improving communications.

Test Directions

Each test should have clearly written directions. These directions should tell the trainee what to do, how to do it, and how to record the responses. General directions should be given for the test, with specific directions given for each section, subpart, and item format. Though the instructor should orally present the directions prior to the start of the test, the written directions should be clear enough to enable the trainees to complete the test without any further instructions. The trainees should be given time to read the instructions and ask questions before the test is started.

Questions that require mathematical calculations pose a unique problem. Suppose a trainee performs a complex equation using a calculator in one step, while the answer key breaks down the calculation into individual steps. The resulting answer will be different from the answer provided in the answer key, since the answer key will break the answer down into individual steps. Each step is then calculated separately and rounded to a significant digit. Rounding of answers (or individual step answers) can cause an otherwise correct answer to be marked as wrong, because the answer key specifies a discrete number. Therefore, precision or accuracy of answers needs to be addressed in the test directions and in the answer key.

Inform the trainees that they may ask questions during the test. Avoid giving individualized assistance by providing any clarifying information from individually asked questions to the entire group.

Trainees should be told the value of test items and how they will be scored. The trainee should know whether partial credit will be given, what degree of precision is required, whether units are required (such as psi, ohms, rem), and for calculations, if work need be shown. Time limits should be stated.

When developing the instructions, keep them clear and concise. Make important points stand out by using a different size type, placing the type in bold, or by underlining. Have an independent review done of the directions to check for inconsistencies or potential

misunderstandings. Consider including sample items with the directions when introducing difficult or unusual item formats. Clear directions will help maintain the reliability and validity of the test. Appendix A provides an example of test directions.

Test Monitoring

Effective test monitoring will ensure that everyone has the same opportunity to understand and answer the questions properly. It is important that the test results provide an accurate indication of a trainee's performance.

Training procedures should provide definitive guidance for test monitoring. A clear policy on academic honesty should be established at the beginning of any training program and should be enforced throughout the program. The single best method is to observe trainees carefully during testing. Some training department procedures require that each trainee sign an affidavit, usually on the test cover sheet, stating the work is the individual's own. This has some deterrent value; however, it should not be allowed to replace other useful methods. These include spacing trainees during testing, using multiple test forms, and revising the test for each session.

4.6 Oral Test Administration

When oral examinations (as opposed to oral questioning) are used, the test questions should be developed prior to administration. The acceptable answers should be recorded in advance along with the applicable references and bases for the questions. This is called pre-scripting and is done to ensure the examination is relevant and valid and provides for consistent tests. The trainee responses should be recorded for evaluation and documentation. The basic procedures for oral examination development are not significantly different from those applicable to written tests. However, the procedure for administering an oral examination has certain key considerations that should be followed.

The number of persons present during an examination should be limited to ensure test integrity and to minimize distractions to the trainees. If a task is performed as part of the examination, a qualified person should be present. Other trainees should not be allowed to witness an oral examination. Oral examinations are not to be used as training vehicles for future trainees. Other instructors may be present either to witness the oral exam as part of their training, or to audit the performance of the instructor administering the test. Others may be allowed to observe oral examinations if (a) the instructor approves the request to observe the test, and (b) the trainee does not object to the observer's presence.

An instructor should brief the trainee prior to beginning the oral examination. Appendix B contains a sample checklist that can be used to assist the instructor when conducting this briefing.

While administering the oral examination, the instructor should allow and encourage the trainee to draw diagrams, flow paths, or other visual representations as appropriate. This allows the trainee to better express him or herself when providing answers or explanations to the instructor. These drawings should be kept with the test documentation. Trainees should be encouraged to use facility forms, schedules, procedures, etc., to answer the questions. The supporting material should be retained by the instructor to provide additional documentation to support a pass or fail determination. The instructor should take sufficient notes during the test to facilitate the thorough documentation of trainee strengths and weaknesses. The instructor should be able to cross reference every comment to a specific subject area question.

The instructor should review and become familiar with the examination material. Prior to the administration of the oral examination the instructor should review any scenario questions with other instructors and discuss the required procedures and special circumstances, etc., related to the scenarios.

The instructor should minimize conversation during the examination. Limit discussions with the trainee during the test to maintain some degree of formality and to avoid distracting the trainee.

4.7 Scoring the Test

Test scoring methods will vary, depending on the purpose of the test. The most common methods are self scoring, hand scoring, machine scoring, and unstructured test scoring.

Self-scoring is often used for tests where the results will not be collected by the instructor. These tests are primarily self-instructional and inform trainees of their current abilities. Self-scoring is also useful for personality, interest, or career planning inventories. Answers can be provided at the end of the test, or a variety of techniques can be used to disclose the correct responses. A variation on self-scoring is to have trainees exchange papers and score them in class. This saves the instructor time and provides immediate feedback for both the instructor and trainee.

Hand-scoring is the most common scoring technique. Usually a scoring key is created on a strip of paper and placed next to the test form, or a blank test form is completed with the correct answers. For multiple choice test items, separate answer sheets can be used. An answer key can then be created by punching out the correct answers. The resulting overlay allows rapid scoring. The overlay should be made of a transparent

material (such as an overhead transparency) so the instructor can easily detect omitted or multiple responses.

When a large number of structured response tests are to be scored, machine-scoring may be useful. In addition to saving time, the ability to enter the results directly into a computer test data base provides many other benefits. Trainee records can be updated, test analysis data can be automatically computed to aid in test refinement and program evaluation, and reports and records can be produced easily once the initial programming is complete.

Many tests are unstructured response format. These tests cannot be machine scored, but should be reviewed individually by the instructor; thus, scoring unstructured response questions consumes a great deal of time and poses some unique challenges. It takes diligence on the part of the instructor to prevent these test items from becoming subjective test items. To minimize the subjectivity in scoring any unstructured response items, several guidelines should be followed.

The instructor should compare the answer key to several of the trainees' responses to a question. Some trainees may take a different approach from what the answer key anticipated and still be correct. If so, an alternate correct answer will need to be added to the answer key. If the key is changed, all tests should then be re-graded using the revised standard.

Periodically review the answer key. It is easy for an instructor's standard to change after several tests are scored. Reviewing the answer key will help protect against distraction from the standard. Also, by occasionally reviewing those items scored earlier, the instructor can confirm that the standards are being applied consistently. Even when applying these measures, some inconsistency is inevitable. One problem is how an item response is graded following several good or several poor responses. The tendency is to score the item low if it follows several high scores, or to score the item high if it follows several low ones. Shuffling the tests between reviews of test questions, while not eliminating the problem, allows these effects to be offset by random sequencing.

Score each item separately. Each test item should be scored for all tests before the next item is scored. Scoring one item at a time allows the instructor to concentrate on just one standard. This increases consistency when assigning points or categorizing items.

Avoid interruptions when scoring responses. The bias an instructor has toward an essay item may change. If a bias exists it should be consistently applied to the responses of all trainees. (For example, an instructor may be irritated one afternoon and calm the next morning) By scoring all response sets at once, if a bias exists, its effects on trainee scores will be consistent.

Provide comments and make corrections on the actual test. A trainee who does not receive full credit for an answer will want to know why. Appropriate comments can explain the score received. If trainees are to learn from their mistakes, they should be told what errors were made and how to correct them. Another value in providing comments is the ability to tally the various comments and analyze the results from test item improvement.

5. PERFORMANCE TESTS

Performance tests measure task performance in the job environment and serve as a mechanism for determining task qualification in the facility. A performance test consistently and systematically evaluates the ability of the trainee to perform a task. Asking trainees to describe proper welding techniques is not a performance test; asking trainees to make a proper weld is. The performance test is not a training instrument; rather it is a testing tool that ensures consistent performance evaluations. A performance test should test both the knowledge and practical requirements that were derived during the analysis of the task.

The steps of a performance test come from the elements of a task. The Training Evaluation Standard (TES) provides the basis for the development of objective-based training materials, and maintains the consistency in the testing of trainee performance. The TES identifies the elements (procedural steps), knowledge, and skills necessary to perform the task. It also identifies the initiating cue that prompts or signals the trainee to begin the task, identifies the terminal and enabling objectives, the conditions under which actions occur, and establishes standards that measure satisfactory performance of the elements, thus the task. A more detailed description on the use of a TES is available in *DOE Training Program Handbook: A Systematic Approach to Training.*

5.1 Developing Performance Tests

Developing performance tests involves identifying economic and other limitations, determining the best instructional and testing methods, and constructing a test that provides the most effective measurement of the task. The task statement, the TES, and the references should be used when developing a performance test. The task statement identifies the task to be evaluated by the performance test. The TES identifies the elements of the task and other supporting information needed for competent performance of each element of the task. The references identified in the TES should be available to the developer when writing performance tests. The developer may choose to make provisions for references, tools, and equipment that are supplied at the time of the trainee test or direct the trainee to gather these resources as part of starting the test.

The following steps should be performed when developing a performance test. The test development process should:

- Determine the testing limitations
- Determine the elements to be tested
- Determine the conditions and the standards

- Determine the method of accomplishment
- Construct the performance test
- Determine the scoring procedures
- Pilot the performance test
- Approve the performance test.

Determine Testing Limitations

The first step in developing a performance test is to review the task and determine the potential testing limitations. Testing limitations are those factors that can have an impact on the development or the conduct of a performance test. These may include availability of time, work force, equipment, and resources. If performance of a task would require more time than is reasonable, the performance test should be developed using only the critical task elements. Work force availability can also impose limitations on task performance. These constraints occur when more than one individual is required for task performance.

Situations occur when equipment or facilities will not be available to support the test. Cost can also affect performance tests. The cost of performance test administration and its effect on consumable repair parts should be kept within reasonable limits. Many infrequently performed tasks cannot be performed for training or testing purposes in the job environment. Safety is another factor to consider. If the testing of certain tasks would impose unreasonable demands on the personnel, facility or equipment, test those tasks using simulation as the method of accomplishment.

The items to be considered when determining the need for simulation or some other method of accomplishment in performance testing are summarized below.

Downtime	Effect of task performance on the equipment readiness and efficiency
Damage	Potential damage to plant equipment and personnel
Cost	Cost of using plant personnel, equipment, and materials

If any limitations result in a change of content in an established test, management, administrative, and instructional approval should be required for the change.

Determine Elements to be Tested

The elements of the task represent an important design consideration. The developer should determine which elements can be tested realistically and should focus on

elements that have the greatest number of skill and knowledge statements. When limitations make performing the task during the performance test unrealistic, task elements should be examined. Elements that include important decision points are predictors of successful performance of the task. If they can be tested realistically, they should be included in the performance test.

A **critical** task element (C) is defined as any element of the task that is deemed crucial to the satisfactory performance of the task. Task elements such as removing interference, obtaining the procedure, and cleaning the job site are typically **non-critical** task elements (NC). Non-critical elements are generally administrative controls and tend to be generic to other tasks. The critical/non-critical designation becomes important in the scoring and evaluation criteria. To determine if an element of a task is critical, consider the following guidelines. An element may be critical if its omission or improper execution:

- Causes or could cause damage to any system or component to the extent that it prevents the system or component from being immediately available for its intended purpose
- Causes or could cause a serious injury or hazard
- Results in incomplete task performance
- Violates security
- Results in an out-of-tolerance condition or measurement which prevents the equipment from meeting facility procedures or specifications
- Violates a standard maintenance procedure such as improper use of test equipment or hand tools, etc. (this does not include performing procedure steps out of sequence)
- Causes excessive delays attributable to insufficient job knowledge or improper planning although the task was successfully performed
- Results in delay(s) due to unnecessary troubleshooting, removal or replacement of components, or rejection of serviceable equipment.

There are some task steps that must be performed in the proper sequence. These should be identified on the checklist for the instructor and the trainee. These steps can be marked with an "S" to indicate they must be performed in sequence.

Determine Conditions and Standards

After testing limitations and element designations have been determined, identify the conditions and standards needed for task performance. Ideally, the test developer

should duplicate the cues, conditions, and standards of the actual task; however, some compromise may be necessary. For tasks with multiple conditions and branching decisions, multiple performance tests may have to be developed.

Conditions are prerequisite requirements that exist for successful task performance. Conditions define facility conditions and include information and resources available to the trainee during task performance. If limitations prevent using all conditions identified in the TES, a sample should be used that best assesses the ability of the trainee to perform the task under actual conditions. Task conditions may require modification if the task cannot be performed under actual conditions. For instance, conditions could include high radiation areas and other environmental concerns.

Performance tests include standards of measurement that are applied consistently in evaluation of task performance. Standards may relate to the process, the product of performance, or a combination of both. Process standards are step-by-step procedures that must be followed, usually without deviation. Product standards prescribe output (the product of performance) and criteria for judging acceptability of the performance (i.e., surface machined to a tolerance of ±.002").

Task standards should be transferred directly from the TES to the performance test whenever possible. However, limitations in the testing environment may require a best approximation of the job standard used during the performance test. Typically the conditions and standards for the elements of a task are implied in the conditions and standards of the entire task. However, if an element has a unique condition and/or standard that is not implied, then it should be stated with that element.

Determine Method of Accomplishment

Each task that is tested should have a designated method of accomplishment (MOA), or level of performance, which dictates how the trainee is to demonstrate the task to the instructor. The MOA is identified for a task and should be identified for the individual task elements so that each trainee is tested in the same manner. There are four possible methods of accomplishment.

P Perform the specified task using approved procedures and observing all applicable safety and administrative requirements. This includes a thorough discussion (usually prior to performing the task) addressing safety implications, elements involved, the effects on associated equipment or systems, and abnormal situations which may arise while performing the task. This method of accomplishment is the most desirable level for performance testing.

S Simulate performance of the specific task. Using approved procedures, "walk through" the task and simulate all actual manipulations (valves, switches, tools, etc.) an employee would perform. Describe applicable safety and administrative requirements and the parameters (meter readings, charts, measurements, etc.) an employee would observe/monitor during actual performance of the task. Conduct the same discussion as required for a perform signature.

O Observe an individual performing the specified task. Conduct the same discussion as required for a perform signature.

D Discuss the specified task using applicable procedures, piping and instrumentation drawings, blueprints, etc., including the discussion as required for a perform. Demonstrate knowledge of the task by describing the manipulations required and the parameters that may be expected to change. This method of accomplishment is the least desirable for performance testing.

Simulate, observe, and discuss should be used only when perform is not feasible, such as in a high radiation area. The trainee should always demonstrate each of the critical task elements by the designated method to successfully demonstrate the task. For example, if the task MOA is "P," the trainee must actually perform each element designated as "critical." The trainee cannot simulate nor discuss those items. The non-critical elements could have a discuss MOA designation to save testing time and allow concentration on the critical items. Non-critical elements need not be included if focus is required on the critical elements to save time.

Construct the Performance Test

Based on the previous information the performance test can be constructed. A performance test typically consists of major items which include:

- A performance learning objective (task statement) indicating the action and the object
- Condition(s) under which the action is to be accomplished
- Standard(s) against which performance is to be measured
- References
- Method of accomplishment (perform, simulate, observe, or discuss)
- Elements (at least critical elements, and non-critical if desired) to be accomplished with the MOA and references indicated

- Knowledge requirements which consist of the cognitive items supportive of the practical requirements
- Practical requirements which consist of the task elements and their related standards.

The questions used for the knowledge requirements should be placed within the evaluation standard to indicate when they are to be asked. Directions should require the instructor to read the questions exactly as written. Space should also be provided to record the trainee's response (if the correct response is not given). The correct answer should always be included with the question.

Additional information from the TES may be included in the performance test such as identifying the task's work group and other information as appropriate. Appendix C is an example of a checklist that can be used when constructing a performance test.

The performance test should not be developed verbatim from a procedure. It should summarize the procedure and be designed to evaluate critical aspects of a particular task. If a task requires specific values such as torque and tolerance, they should be stated in the standard for the task or element. Hold points should be inserted at desired locations in the performance test to allow the instructor to grade the trainee's performance of the previous steps.

The performance test package should consist of an administrative section, instructions to both the instructor and the trainee, a guide for the instructor to use for scoring, a trainee guide, an optional data sheet for trainee use, and a section used for documentation (e.g., a check-off list). Appendices E and F are two examples of performance tests.

Develop Scoring Procedures

The developer should create an evaluation instrument that the instructor can use to accurately measure the trainee's performance of each step of the performance test. The evaluation instrument should accurately measure the trainee's ability to demonstrate the task.

When the performance test is constructed, scoring procedures are developed. A detailed, step-by-step description of required performance provides an effective scoring procedure for some tasks. Action steps or elements required in the performance test usually are prepared in checklist form, and the trainee is required to follow each step without deviation. For other tasks, the product of performance (i.e., a tangible result) should be measured. In developing a scoring procedure for this type of performance

test, scorable characteristics need be defined to distinguish clearly between satisfactory and unsatisfactory performance.

Scoring methods should adhere to administrative and instructional guidelines and reflect the evaluation standards. If an evaluation standard is "without error" a yes/no checklist should be used. If the standard implies some range of acceptable performance, a rating scale may be used. However, rating scales introduce greater subjectivity and are more difficult to use, to interpret, and to back up than a yes/no checklist. If sequence is important, identify this on the performance test and provide proper scoring guidance.

Establish cutoff scores to meet the performance standards. Percentages are the least preferred method. Failure of a performance test should be determined by the failure of any critical step or the failure to follow required sequences. The cutoff score for any behavior should be based on a level of performance that is achievable by the average person.

All methods of scoring should be consistent with policy, procedures, specifications, and needs. The method should also adhere to instructional guidelines, such as testing the objectives, and should clearly distinguish between satisfactory and unsatisfactory performance.

Piloting the Performance Test

The purpose of piloting a performance test is to ensure that the "bugs" have been worked out of the test. The pilot should be conducted under the conditions required for actual job performance or the same conditions under which the trainee will be tested.

The pilot should have two evaluators monitor the performance of a single individual. This should be a simultaneous but independent evaluation. If the scores (sat/unsat) are different for any of the steps, a reliability problem exists. When conducting the pilot the evaluators should look for problems or deficiencies such as:

- Questions asked by the trainee
- Equipment requirements
- The ability of the trainee to perform the task
- Unclear directions to the trainee
- Unusual conditions or problems beyond your control that affect the outcome of the test
- The effectiveness of the scoring method used

- Time considerations.

Approve the Performance Test

After the performance test has been piloted, reviewed, and corrected from feedback, it should be approved. The test package should be signed and dated by both facility and training representatives. Appendix D is an example of a review checklist for a performance test.

5.2 Test Administration

Test administration has an important effect on the usefulness of test results and requires control. The instructor should ensure that a suitable environment is established, clear test directions are given, and proper supervision is present for the entire test.

Establish Environment

Effective testing requires that the physical qualities of the test environment and setting that the trainee performs within are satisfactory. High noise levels, poor lighting, lack of ventilation, excessive heat or cold, adverse safety conditions, and frequent interruptions will lower trainee test performance. Prior to the performance test, the instructor should ensure that the conditions of the test location are adequate.

Test Directions

Prior to conducting a performance test the instructor should provide the trainee with directions and an overview of the performance testing process. These directions should provide the trainee with clear and complete instructions as to what the trainee will be allowed to do, and when the instructor will allow the trainee to do it. The instructor should explain under what circumstances he/she will stop the trainee if conditions such as safety of personnel or equipment arise.

Conducting the Performance Test

The completion of the task is not the only indicator of the competence level of the trainee. It is important to observe the methodology as well as the outcome of the performance test. Some typical questions that the instructor should consider when observing a performance test include:

- Were the tools used correctly and in the proper sequence
- Were the necessary reference materials obtained
- Were non-critical steps performed in the proper order
- Was the trainee confused by any portion of the performance test

- Was the equipment manipulated in a deliberate and timely manner
- Was the trainee aware of equipment status (e.g., did he/she recognize when a pump was running or when a valve was open)
- Were safety rules observed when performing the task?

Complete testing of the knowledge and skills requires the instructor to question the trainee during the performance test; however, the instructor should not ask distracting questions. All questions should be related to the task. The instructor may ask the trainee to "talk through" the task as he/she performs it. This technique reduces the number of questions the instructor needs to ask and allows the instructor to stop the trainee before he/she makes a serious mistake. The questions may be written in the evaluation standard (preferred method) or generated by the instructor during the performance test.

During the conduct of a performance test the instructor is also a safety monitor in addition to his/her role as evaluator. The instructor has the responsibility of stopping the performance test whenever personnel injury or equipment damage can occur, public health or safety is affected, or the trainee deviates from an approved procedure.

Evaluating the Performance Test

Scoring methods should be identified and should be closely related to the evaluation standards. Trainees should be evaluated on how closely their performance meets the standards. Some rating method examples are pass/fail, sat/unsat, yes/no, and 80% correct.

Debriefing the Trainee

At the completion of a performance test the instructor and the trainee should conduct a detailed review of the trainee's performance. The instructor should tell the trainee if he/she passed or failed the performance test. The review should be conducted immediately while the events are fresh in the mind of both the instructor and the trainee. The instructor is responsible to record the results accurately before, during, and after the performance test. Accurate recording of results allows the testing process to be evaluated, ensures fair grading, and allows for monitoring the results to ensure reliability.

6. TEST ANALYSIS

Because tests are used to qualify trainees to do a job or task, it is important that they are developed properly. If tests are constructed systematically and administered correctly, they will have a high degree of reliability. The quality and effectiveness of tests should be continuously monitored and improved where necessary. Analysis of test results provides important input to the quality and effectiveness of tests. Whereas most instructors and test developers are not required to perform complicated statistical analyses, an understanding of some basic concepts is beneficial in interpreting and refining the testing process.

6.1 Reliability

Reliability is functionally defined as the consistency between two separate measurements of the same thing. If a test gives perfectly consistent results, it would be perfectly reliable. Reliability is generally not a problem with performance tests as long as conditions in the evaluation situation remain constant. Reliability can be a problem with written tests because test item construction can be difficult. Reliability can be affected by ambiguous test items, multiple correct answers, typographic errors, adverse testing conditions, interruptions, limited time, and complicated answer sheets. Trainee readiness and scoring errors also affect test reliability.

The following examples illustrate how reliability or unreliability may be indicated as tests are analyzed.

> Example: Ten trainees were given test A on Monday and then again on Tuesday. Assuming that nobody forgot anything overnight, the Tuesday test results should be exactly the same as the Monday test results if test A is reliable.

Any significant difference would indicate test unreliability since nothing changed from Monday to Tuesday. This is a form of test-retest reliability. The time period for this type of reliability is variable. Longer time periods generally result in greater differences in test results, but long time periods can determine the long-term stability of the test.

> Example: Ten trainees took a test and 9 of them missed question #5. Question #4 was missed by nobody but was testing an item very similar to that covered by question #5.

Question #5 may be unreliable due to poor wording, unclear answers, a typographic error that makes a wrong answer look correct, etc. This is a form of alternate question reliability.

Example: 8 of 10 trainees who missed question #7 chose answer (b). Does answer (b) look too similar to the correct answer? Does the lesson plan support the correct answer?

The above example could indicate the use of a method of testing known as key word and tricky phrase testing. This type of testing causes the trainee to memorize and recall only key words and tricky phrases to pass the test instead of requiring the trainee to learn the material; thus it is a poor method to use.

Test items with poor reliability are easy to recognize. If trainees that are equal in knowledge or ability have widely varying test scores, the test or test item may be unreliable. Or, if the same trainee is tested twice on the same test or test item within a short period of time and passes once and fails the next time, the test or test item may be unreliable. In both of these cases the reliability should be questioned and the test or test item should be carefully evaluated.

6.2 Validity

A valid test measures exactly what it was intended to measure. A test can be reliable but not valid, or valid but not reliable. A paper and pencil test can be reliable in measuring knowledge of certain welding fundamentals, but not valid for measuring welding skill. Establishing the validity of tests can be a complicated and time consuming process. Validity can be improved by:

- Ensuring a good analysis of tasks has been conducted
- Ensuring that knowledge and skill requirements have been identified
- Ensuring that learning objectives for both knowledge and skills are based on task requirements
- Identifying type of performance dictated by objectives (cognitive, psychomotor, affective)
- Ensuring action verbs used in objectives measure what they were intended to measure
- Designing test specifications to ensure that objectives are covered adequately
- Discussing the test with SMEs, supervisors, and training specialists
- Piloting the test or sample test items with SMEs and trainees
- Comparing test results to actual job performance
- Ensuring that the test and test items are changed to be consistent with revised job requirements

Content Validity

Content validity is the simplest method to assess whether a test is valid. Establish content validity by comparing the test items to the learning objectives. No statistical calculations are used to establish content validity. If subject matter experts agree that the test items measure their respective learning objectives, the test can be considered valid. The usefulness of content validity is subject to the quality of the analysis and the subsequent learning objectives as well as the thoroughness of the SME review of the test items.

Concurrent Validity

Concurrent validity of a test is when one test compares favorably with another, already validated test. If there is already a valid measure (i.e., nationally recognized entrance exam) of what is to be tested, determine the degree of association between the results of the pre-established test and the test to be validated. To the extent that they are related, there is an established level of concurrent validity. A statistical analysis is required to establish a level of concurrent validity. Information on statistical analysis to determine concurrent validity can be found in several commercially available textbooks on statistics.

Predictive Validity

Predictive validity is when trainee scores on one test can be used to predict success on a second test after a given time interval. Establishing predictive validity is accomplished in a similar manner as establishing concurrent validity. Statistical analysis is used to determine predictive validity as long as both tests are scored on a pass or fail basis and the tests are separated by a substantial period of time.